Chain-of-Verification Reduces Hallucination in Large Language Models

Year: 2023 | *Citations: 308* | *Authors: S. Dhuliawala, M. Komeili, Jing Xu*

Abstract

Generation of plausible yet incorrect factual information, termed hallucination, is an unsolved issue in large language models. We study the ability of language models to deliberate on the responses they give in order to correct their mistakes. We develop the Chain-of-Verification (CoVe) method whereby the model first (i) drafts an initial response; then (ii) plans verification questions to fact-check its draft; (iii) answers those questions independently so the answers are not biased by other responses; and (iv) generates its final verified response. In experiments, we show CoVe decreases hallucinations across a variety of tasks, from list-based questions from Wikidata, closed book MultiSpanQA and longform text generation.